ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

ROYAL STATISTICAL SOCIETY RESPONSE TO GOLDACRE REVIEW INTO USE OF HEALTH DATA FOR RESEARCH AND ANALYSIS

Key points

The Royal Statistical Society (RSS) welcomes this important initiative. Health data is an important public good. We have a responsibility to make effective, responsible use of it, for the benefit of patients, the population as a whole, and the advancement of knowledge.

There are important practical steps that the government should take to make the most of health data:

- Invest in mapping the health data system, to identify valuable data sets that could be shared and gaps where investment in data is necessary
- UKRI should expand its funding for data sets and tools, backing projects that (i) bring together new data sets, and (ii) clean existing data sets and protect confidentiality of sensitive information in these so that they can be released to researchers in a responsible way. In doing so, it should partner with funders like Wellcome Trust and the Health Foundation.
- Focus on making these data sets easily accessible. UKRI should fund “sandboxes” (in both academia and industry) and incubators for startups to work with new data sets, perhaps focused on important health-related missions – such as preparedness for future pandemics.

Protecting the privacy of individuals contributing their information to these databases and in general building and maintaining public trust in this process are essential to making these things work. There should be a ‘context-sensitive’ data governance & ethics structure put in place, which allows flexibility dependent on the nature of data usage being considered. This may involve applying statistical disclosure control methods to protecting confidentiality in the data while still retaining the value of the original data, eg creating representative synthetic versions of the original data. Clear and transparent communication of the methods used will be crucial in ensuring the data are still seen to be of high value and confidence in providing information to these databases is not diminished – there should be an open dialogue with the public at large, to create transparency and build trust.

Finally, all this must be supported by investment in skills. People who work with data in the NHS are an often-overlooked group. They deserve greater professional recognition and skills. Accreditation systems, like the RSS’s Data Analyst accreditation, can help provide this.

Addressing the Terms of Reference

Rather than providing single answers to the terms of reference we have engaged with the Society's members to identify a range of considerations and proposals that might helpfully inform the review in each of these areas.

1. How do we facilitate access to NHS data by researchers, commissioners, and innovators, while preserving patient privacy?

- The general aim should be to identify efficient and innovative methods of data anonymisation or de-identification, which balance protecting privacy and removing personally identifiable information with the utility of the data provided to allow a wide range of novel secondary analyses. Synthetic data is increasingly using an approach with the potential to achieve these goals. The basic idea is that a statistical model is used to simulate values to replace original values in the data. Provided a plausible model is used statistical properties in the synthetic data should be similar to those present in the original data, resulting in better data utility. Also, as the data now contain synthetic values, confidentiality will have been protected to an extent.
- Local benefit should be prioritised over a one-size-fits-all national approach to health data science. That is, there must be enough of an incentive in each UK healthcare region/area for them to manage healthcare outcomes and quality better, from the way their data is used, rather than just asking them to contribute to a national scheme. Central organisations (eg NHS-Digital) would still have an important role to play, and would be needed to set nationally consistent standards and frameworks for data curation and linkage, so that we get the best of both worlds – local value, as well as consistent nation-scale analytics
- Create sandbox tech environments, with privacy-controlled data feeds, to enable space for smaller companies to innovate and compete on the data – eg, to enable startup incubators.
- Consider synthetic data approaches to enabling safe and broader scientific and administrative use of healthcare data, including differential privacy techniques
- Consider an auto opt-in approach to sharing of electronic medical records (by individuals, practices). Specific incentives to share may also be worth considering.
- Ensure UKRI funding (through the Research Councils et al) is directed to the right areas for future preparedness – eg, big data studies of immunomics and virus dynamics, in the wake of Covid-19.
- The context in which the data would be used also bears consideration and it may be worth considering differentiating along several lines: public health emergency contexts; public health non-emergencies; individual health care focused research, and individual health care focused practise of medicine. In some of these contexts there may be an argument from the public good to vary how data is accessed.
- It is worth considering the multiple stakeholders around issues of what (big) data is, and how it should be classified: as an asset, a resource, private property, a public or common good etc. Flowing from this, can or should we speak of owning or controlling data; or are concepts such as stewardship appropriate?
- Who 'owns' NHS data in a pandemic (and in all other contexts); who should control it; how /who ensures the appropriate public health governance? Analysing the appropriate framework for facilitating access to NHS data needs particular attention to commercial sector NHS data access, differentiating between public health, individual health, research, practice of medicine, emergency, and non-emergency.
- Currently there is a shortage of agreed substantive and procedural data governance and ethics standards in general – and it might be helpful to consider how to address this. Work is urgently needed to expand public health global ethics and considerations of health as global public good (in an economic and moral sense). The work of the Nuffield Council on Bioethics is an important resource when looking at all these matters. Accepted bodies of thought that interface with data such as the Caldicott Principles need to both be covered by a data review, and also critically reviewed in the light of our 'data innovation era' (earmarked by both vast quantities of data, the technology and skills to

analyse it, and the experiences of our pandemic age). Originally six principles, an eighth Caldicott Principle was added in December 2020, namely the principle that duty to share information can be as important as the duty to protect patient confidentiality.

2. What types of technical platforms, trusted research environments, and data flows are the most efficient, and safe, for which common analytic tasks?

- Remote analysis environments (password / VPN etc. protected) add an additional level of security but at the cost of efficiency and flexibility of analysis for some types of analysis (eg analyses which combine evidence, particularly if datasets are held in different environments and cannot be combined.
- Statistical disclosure control methods (like synthetic data mentioned above) offer the ability to produce a less risky version that can be released more freely with less confidentiality concerns. This would allow users the ability to use the data more flexibly. It could also provide users with a “test data set” allowing them to become familiar with the type of data to expect while they wait to gain access to the original data this reducing the amount of users’ time and effort.
- A related idea to remote access is that of verification servers. This is where users submit queries about the data and they receive answers to their queries provided there are no confidentiality concerns. This can tie in with synthetic data where users submit the results of their analysis conducted on the synthetic data and a verification server returns information about how close the analysis is to the corresponding analysis performed on the original data. Again, provided there are no confidentiality concerns.
- Consider open standards and accreditation for software, to ensure a minimum quality level. For instance, code used to guide UK government policy at the start of the pandemic was many 1000s of lines of old, undocumented C code, originally developed for another purpose. It should not take too much time and effort to turn that into something robust and ‘production ready’, yet there didn’t seem to be a clear route to do that.
- It will help improve efficiency and safety if code is made open for public and expert scrutiny.
- Similar for modelling and data visualisation: consider best practice and standards, eg things like insisting on sensitivity analysis and error bands for modelling, and do’s and don’ts for visual communication of data findings to the public
- Examine emerging techniques in the AI/ML domain for potential beneficial reuse in the health data space. For instance, use of GANs (Generative Adversarial Networks) for creating realistic “fake” medical records.

3. How do we overcome the technical and cultural barriers to achieving this goal, and how can they be rapidly overcome?

- Here there need to be substantial investment in computing resources in order to 1) Store the data securely and 2) Implement sometimes complex computational methods to protect the data. Investment will also need to be made in developing novel statistical methods to anonymise and protect the types of data being considered, eg using synthetic data methods. The type of data will drive the most appropriate statistical method to use, eg are the data longitudinal or geographic in nature, do they comprise household or just individuals?
- Culturally there needs to be engagement with the stakeholders about the methods being employed. For the data holders this will involve assuring them any data released will be safe. This will naturally involve eliciting what confidentiality concerns they have from the data and what features are most in need of confidentiality protection. For users of the data there needs to be engagement to ensure that any released data is still of high utility to them. This will involve finding out from users what purpose they would like the data for.
- The RSS recently organised a scientific meeting on data confidentiality in medical data that led to some interesting discussions and highlighted some of the key issues in this area. Holding more

meetings like this would facilitate discussion around this area and help build a consensus on how best to proceed.

4. Where (with appropriate sensitivity) have current approaches been successful, and where have they struggled?

- Currently in the UK there has been relatively little development in methods to release confidentiality data on a practical setting and this is something that is greatly needed. More work on this has been done in the US, eg the US Census Bureau have a synthetic version of the Survey of Income and Program Participation. In the UK the Office for National Statistics does have active programmes of research on synthetic data, both within the Statistical Disclosure Control group and the ONS Data Science Campus (together with Dr. Robin Mitra at Cardiff University) and a key goal is to be able to produce practical versions of synthetic data with promising early results at this stage.
- There are many aspects of the Covid-19 pandemic that direct us to urgently re-examine attitudes and approaches toward data sharing and data access including the phenomena of state-led ‘vaccine nationalism’, state-led research and supply chain nationalism, with governments acting to ensure that their own country has priority access to health care. Can one justify or even require the adoption of ‘health data protectionism’ in some circumstances, although the risks attached to this approach are clear?
- Research and data nationalism relies on the cooperation of the involved pharmaceutical companies: can /should they remain politically, ethically neutral in their business activities? At what point must action be taken at state level against data profiteering, etc.? Could a modern re-working of the state’s powers to undertake compulsory land purchases be needed in the world of health data?

5. How do we avoid unhelpful monopolies being asserted over data access for analysis?

- Follow steps used in other areas that avoid unhelpful monopolies.
- The use of the term “unhelpful monopolies’ needs careful deconstruction; various points touching upon this wording have been offered above. For example, discussions around whether ‘ownership’ or common-ownership of data is an appropriate concept are inextricably connected to IP/patenting. Data has a particular set of properties that differentiates it from physical assets; datasets can be duplicated at near-zero cost, used in multiple ways by different people without diminishing their value. The value increases as they are combined with other datasets; it is not a finite resource; the use of a dataset for one purpose does not inhibit the use of the same data for social or non-commercial purposes. Do we need a new lexicon regarding data and ‘monopolies’? A further interrelated issue is the question of ascertaining the value of data and if (when) compensation (or benefit sharing) should be paid for access to data. Such discussions have usually focused on the questions that arise if state agencies provide public data to a commercial enterprise.
- Regarding NHS data: a fundamental question is whether the state should be seen as holding a ‘helpful monopoly right’ to ‘our NHS data’? One speaks of ‘our NHS‘; should we act based on ’our health data’? If yes, how would this look in reality; how far should the power of the state over ‘our’ health data be presumed; how far should consent be seen as being waived or delegated to state agencies? How much power should we give to the state: in an emergency; regarding data on a public health level; regarding individual level data?
- If NHS can have a helpful data monopoly, the question is under what terms and conditions should/could/can the NHS grant third party licences to use its data in order to bring health benefits.
- How about commercial R&D patents that have been built on access to data RWD including ‘our’ data? Are the benefits that arise from new medicines sufficient pay-back to society?

6. What are the right responsibilities and expectations on open and transparent sharing of data and code for arm's length bodies, clinicians, researchers, research funders, electronic health records and other software vendors, providers of medical services, and innovators? And how do we ensure these are met?

- Completely 'open' sharing may not be suitable for all contexts, but for transparency, it is the responsibility of the data provider to clearly explain and justify why certain data cannot be shared.
- The Office for National Statistics explains what confidentiality guarantees they provide with the data they collect on their website. These could be used as a starting point when beginning to address this question and perhaps be discussed further through meetings with key stakeholders to develop appropriate guidance for everyone involved in this process.
- The comments in response to the previous question suggest that there can be no one set of right responsibilities and expectations when sharing NHS (public) data. The work done in particular by the Nuffield Bioethics Council must be referenced when constructing such codes, as well as the considerable body of high level analysis already undertaken.

8. How significantly do the issues of data quality, completeness, and harmonisation across the system affect the range of research uses of the data available from health and social care? Given the current quality issues, what research is the UK optimally placed to support now, and what changes would be needed to optimise our position in the next 3 year?

- There are likely to be a lot of issues with the current quality of the data, eg missing data or measurement error that will cause complication in the analysis of the data and perhaps also lead to inaccurate conclusions being drawn. There are standard measures to address these but these may have varying performance depending on the data being considered and the severity of these problems. So while some analyses could be performed on some data sets there needs to be greater research support for methods to handle these issues in the data from a practical point of view. There also needs to be support for research into methods to protect the confidentiality of the data appropriately, eg developing appropriate synthesis methods here.

9. If data is made available for secondary research, for example to a company developing new treatments, then how can we prove to patients that privacy is preserved, beyond simple reassurance?

- Provide detailed information in relation to the steps that are taken in order to "anonymise" data. Involve patients in the oversight committees which approve steps.
- We can also run various "intruder attack scenarios" eg pretend we are a malicious user of the data (often called an intruder) and try to uncover sensitive confidential information from the released data. If these scenarios result in too high a level of the risk, we can go back and apply more stringent confidentiality protection measures until the risk is below an acceptable level. We can provide users the results of the final round of attack scenarios to reassure them that their privacy is well preserved.
- There is no one set of appropriate instructions for research using secondary data. The extent to which personal privacy is the highest good that can / should / must be protected must vary according to context. A wide stakeholder discourse must be the prequel for devising a set of processes and principles on these matters.

10. **How can data curation best be delivered, cost effectively, to meet these researchers' needs? We will ensure alignment with Science Research and Evidence (SRE) research priorities and Office for Life Sciences (OLS) (including the data curation programme bid).**

- This is where measures like synthetic data (and perhaps other statistical disclosure control methods) offer good value for money. Providing an easily accessible data set, such as a synthetic data set, that is deemed safe for release to users, will save users' time, and thus costs, spent analysing the original data. This is because users can test and develop their methods on the synthetic data reducing the time spent needing to do this on the original data. Further the agency responsible for protecting the original data will need to spend less resources, and thus costs, verifying users' analysis of the original data is safe as much of the users' preliminary analysis will have taken place on the synthetic data already and thus not required to be performed on the original data again.

11. **What can we take from the successes and best practice in data science, commercial, and open source software development communities?**

- Covid has shown the importance of collaborations, some examples of which are: the Covid-19 Therapeutics Accelerator launched in March 2020 by the Gates Foundation, Wellcome, and Mastercard; the Pharmaceutical Research and Manufacturers of America Biopharmaceutical Industry Principles on Beating Coronavirus; Medicines Patent Pool, with important examples of patent rights being renounced in favour of the public good; and OpenSafely.

12. **How do we help the NHS to analyse and use data routinely to improve quality, safety and efficiency?**

- Provide detailed case studies where this is happening so that others can follow. Preferably detailed case studies where issues have arisen so as to avoid others going along similar routes.
- Regular meetings that involve staff from the NHS and the Medical and Pharmaceutical sector more widely, but also bring together academics and staff from the ONS to share best practice could help. The recent RSS event along these lines (mention in response to Q3) resulted in some very fruitful discussion. Events like this, focused on issues experienced by the NHS could be very helpful.

Appendix: Perspective from RSS Data Ethics and Governance section

This appendix provides a more overarching view on the review, from the perspective of RSS members especially concerned with data ethics and governance.

1 General Comments

The focus of the new review is “the more efficient and safe use of health data for research and analysis for the benefit of patients and the healthcare sector.”

This work is undoubtedly important, relevant, and much needed; modern health research, innovation and health care is increasingly data driven and data assisted. A productivity, outcome-based focus on ‘efficient’ data use has a role to play as we all try to react to the Covid lessons-learnt, and act on Michael Marmot ‘s recent comment that “England is Faltering”, with falls in life expectancy being registered for some populations.[1]

However, ‘efficient and safe’ data usage must also be socially and morally acceptable; data use must satisfy governance best practices and comply with the relevant consensual ethics principles; must, inter alia, be fair, just, transparent, accountable, and trustworthy.

The terms of reference do not make clear that this foundational or parallel step will be undertaken when evaluating ‘efficient and safe data use.’

2 Structuring Work on the Use of Health Data

2.1 Introduction

Formulating recommendations to government on the role of data in generating benefits for patients and the healthcare sector must take a stringent and structured account of the complex, dynamic role of data in health; Covid-19 has shown the positive importance of data in both research and non-research healthcare, as well as illustrating that data use trustworthiness must be earned, not assumed. Though it must be noted that a global health pandemic is a very particular situation, not to be lightly used as basis for extrapolation to data use in other settings.

A review of the potential uses of data needs to be highly granular and rigorously analytical in order to do justice to the dynamic and multi-dimensional nature of health (assuming that the WHO aspirational multi-dimensional definition of health as a state of complete physical, mental and social well-being, and not merely the absence of disease or infirmity is being here used). The following paragraphs outline the various parameters, contexts and situations that must be differentiated when conducting a health data review.

The interoperable data needed for research and analysis in order to add benefits for patients and the healthcare sector must cover a huge range of data to be in any way meaningful. Careful process and stakeholder maps must be drawn up for all classes of health data use, encompassing the public, government at all levels, research conducted in academia, PPPs, and the private sector. The stakeholder maps need to have a global component; both data and threats to health have a natural *laissez-passer*.

[1] Health Equity in England: The Marmot Review 10 Years On, 2020.

Although questions (and responses) regarding the ‘more efficient and safe use of health data’ must be set in the context of a particular health care system, the review is advised to look to such material not only on a national, but also international level. Efficiency can be achieved via access to relevant data held in other countries (eg, vaccine trials) but safe use requires that data obtained from other countries have been through the same rigorous ethical checks and balances as would apply in the UK. Such transparency is only possible via international cooperation leading to the development of trusted partnerships.

2.2 Non-Exhaustive Classes and Categories of Health Data Use

2.2.1 The use of the phrase “use of health data for research and analysis” suggests that the review appreciates that a difference must be made between

i. data use in research (necessarily separating clinical and public health research).
ii. data use in clinical practice/analysis outside research.
iii. data use in the mixed/ continuum space linking and mixing research and practice of medicine.
iv. public health non research data use.
v. public health research data use.
vi. data use in regulatory affairs and regarding post marketing surveillance.
vii. data use in data quality control and other quality control.

2.2.2 The structure of a review of health data use must differentiate between:

i. health data research and analysis on an individual, personal data level (i.e. covered by GDPR).
ii. standard public health context.
iii. data used to locate and define a community.
iv. and data then needed to analyse a specific community[2] (such as ethnic minority groups).
v. population level and individual data in an emergency (including looking at the power of the state is issuing pandemic emergency legislation).

2.2.3. Ongoing work in Covid-impacts on different ethnicities has reinforced our appreciation that in addition to physical health status end-points, data on the causes and impacts of health status must be understood and obtained (looking at mental, social, environmental dimension of health). The data review must cover the determinants/causes, pressures, and drivers of health on both individual, community and population levels.

2.2.4 It is furthermore vital that all ethical, legal, and commercial health data governance analysis differentiate between data processing undertaken wholly in a public agency (NHS) setting, and data activities involving the commercial sector. For instance, regarding privacy, public agencies are able to draw upon some level of presumed consent (or waiver of consent) covering some uses of data; commercial sector actors must take another approach.

3 Particularly Challenging Health Data Analysis Tasks

Here are but a few of these:

[2] The responsibilities of using data to define a ‘community‘ (on social, genomic, location, economic, faith-base, heritage criteria) are considerable and must be established as part of a data review.

3.1 As mentioned above, health dimensions and their determinants are furthermore not discrete but interwoven, interacting, and interdependent, with a change in one determinant and dimension affecting and driving the others. The challenges of identifying, linking, and providing appropriate access to a matrix of social, economic, ethnic data that are connected with physical and mental health outcomes are considerable not merely from an ethical point of view, but face problems because of the different governance and legal systems that cover data sharing. While regulatory and advisory bodies have been established[3] to facilitate research use of personal data in the domains of health and of social and economic data, attempts to combine records across these domains face a double hurdle.

3.2 The responsibilities of using data to define a 'community' (on social, genomic, location, economic, faith-base, heritage criteria etc) are considerable and must be established as part of a data review.

3.3. An area needing special data review attention is data and the commercial health R&D sector. Reminiscent of the HIV-AID crisis, the Covid pandemic has drawn attention to the vital role of the commercial sector (acting alone and in collaboration with academia) in research, manufacture, and global supply of ingredients and finished goods. Innovative research methodologies rely increasingly on access to high quality data; the commercial sector is increasingly active in acquiring vast swaths of 'real world data,' proclaiming that access to (or ownership of) health data gives a substantial competitive advantage.

The 'data social contract' between society, the state and industry needs to be re-evaluated and possibly re-negotiated in the light of big data and real world-data, particularly in the face of a global health pandemic. To the extent that commercial research increasingly draws on access to real world data in its R&D work, should revised benefit sharing / patent-derived super-profit-sharing arrangements be considered – particularly in a global health pandemic?

How should government, society and the NHS view providing commercial pharma with NHS data? Should we be look back historically at war-time attitudes towards profiteering?

The work of the NHSx is to be noted on the challenging issue of sharing NHS data with the commercial sector.

The Department of Health and Social Care issued "Code of Conduct for Data-driven Health and Care Technology" (published September 2018, updated 19 February 2019) that "encourages companies to meet a gold-standard set of principles that will protect patient data and make sure only the best technologies are used by the NHS, to bring real benefits to patients."

Attention also deserves to be given to WTO regulations, intellectual property, (patents), and documents such as the Doha Declaration on the TRIPS Agreement that affirms the right of developing countries to take action to protect public health and provide access to medicines for all.[4] Concepts such as compulsory licensing also need to be looked at; can anything be learnt that is applicable to health data? The role of the WHO as data collector

[3] For example, the Confidentiality Advisory Group assists with applications for access to non-consented health records, whereas the National Statistician's Data Ethics Committee and the UKSA Research Accreditation Panel assist with applications for access to non-health data.

[4] The WHO Global Strategy and Plan of Action on Public Health, Innovation and Intellectual Property (GSPA-PHI) and the WHO Roadmap for access to medicines, vaccines and health products 2019-2023: comprehensive support for access to medicines, vaccines and other health products.

and data repository of first resort in a pandemic needs consideration, as must the repercussions of attainable health as human right when looking at access and sharing of health data.

Thus, a review of a “more efficient” use of health data requires an examination of the morals and economics of ‘heath data’ as a global public good and data as source of political and economic power and influence.

4 Statistics Methodology and Complex Medical Data

A meaningful (and not harmful) use of medical data in any situation can only be developed in partnership with statistics expertise; the involvement of the RSS Royal Statistical Society in the medical data review and all follow-up projects is essential.

5 Existing Source Documents

Addressing the questions that make up the “Terms of Reference” must take note of the very rich landscape that has grown up surrounding data governance and data ethics, a landscape populated by many instances, bodies and agencies who have or are issuing important reports, guidance, and framing documents. There are also a large number of projects underway in the complex web of legal, governance, regulatory and ethics fields that embrace the dynamic field of health data:

- National Data Strategy; NHSx data strategy for health and social care.
- Work of Department of Health and Social Care, e.g. “Code of Conduct for Data-driven Health and Care Technology” (published 5 September 2018, updated 19 February 2019.
- DHSC 2021 report “Integration and Innovation: working together to improve health and social care for all” legislative proposals for a Health and Care Bill CP 381 Published 11 February 2021; DHSC Policy paper The future of healthcare: our vision for digital, data and technology in health and care Published 17 October 2018.
- DHSC Consultation outcome report ‘Busting bureaucracy: empowering frontline staff by reducing excess bureaucracy in the health and care system in England’ Updated 24 November 2020.
- NHSx various activities & docs; Health and Care IG Panel; NHSX Data Strategy for Health and Social Care (forthcoming).
- HDR UK data sharing principles; Ada Lovelace documents; Nuffield Bioethics Committee reports.
- Various Royal Societies, various Public Health agencies and professional bodies.